Received: 31 July 2019 | Revised: 17 March 2020 | Accepted: 9 April 2020

DOI: 10.1002/for.2693

SPECIAL ISSUE ARTICLE

WILEY

Modeling of frequency containment reserve prices with econometrics and artificial intelligence

Emil Kraft | **Dogan Keles** | **Wolf Fichtner**

Institute for Industrial Production (IIP), Chair of Energy Economics, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

Correspondence
Emil Kraft, Chair of Energy Economics, Karlsruhe Institute of Technology (KIT), Institute for Industrial Production (IIP), Hertzstraße 16, 76187 Karlsruhe, Germany.
Email: emil.kraft@kit.edu

Funding information
Bundesministerium für Wirtschaft und Energie, Grant/Award Number: 03SIN120

Abstract

The forecasting of prices for electricity balancing reserve power can essentially improve the trading positions of market participants in competitive auctions. Having identified a lack of literature related to forecasting balancing reserve prices, we deploy approaches originating from econometrics and artificial intelligence and set up a forecasting framework based on autoregressive and exogenous factors. We use SARIMAX models as well as neural networks with different structures and forecast based on a rolling one-step forecast with reestimation of the models. It turns out that the naive forecast performs reasonably well but is outperformed by the more advanced models. In addition, neural network approaches outperform the econometric approach in terms of forecast quality, whereas for the further use of the generated models the econometric approach has advantages in terms of explaining price drivers. For the present application, more advanced configurations of the neural networks are not able to further improve the forecasting performance.

KEYWORDS

artificial neural network, balancing reserve, econometrics, electricity price, time series forecasting

1 | INTRODUCTION AND MOTIVATION

Transmission system operators (TSOs) have responsibility for a secure electricity system operation, which includes ensuring a stable grid frequency of 50 hertz within their designated control areas. This is achieved by continuously balancing power feed-in and withdrawal.

To balance frequency perturbations, balancing reserve capacity is deployed by the TSOs. Balancing reserve capacity is characterized by a short reaction time and the ability to increase or decrease the power feed-in quickly upon request. Depending on the response and the activation time, three different qualities are distinguished in continental Europe. The different quality requirements lead to market segments for primary (frequency containment reserve, FCR), secondary (automatic frequency restoration reserve, aFRR), and tertiary (manual frequency restoration reserve, mFRR) balancing reserve power, in which FCR has, at 30 seconds, the shortest activation time. In the past, mainly conventional generation such as nuclear, coal and gas power plants, but also hydropower, were the only providing technologies of balancing reserve power. In recent years, new technologies entered the market and, by today, renewable energies such as biomass, photovoltaics and wind power, but also battery storage, are technically capable of providing balancing reserve. Because market

liberalization TSOs are not allowed to own generation capacity, they procure positive and negative reserve capacities meeting different quality requirements through public tenders. These markets for balancing reserve coexist alongside derivative and spot markets for electricity, enabling additional return opportunities for generators by meeting the respective requirements.

The auctions for FCR take place on a weekly basis each Tuesday at 3 p.m. and are dedicated to the provision of FCR in both a positive and negative direction for the following week. Market participants place a capacity price bid and are compensated according to pay-as-bid pricing.

This paper focuses on forecasting the prices of the largest European FCR market, in which the TSOs of the control zones of Austria, Belgium, France, Germany, the Netherlands, and Switzerland jointly[1] procure roughly 1.4 gigawatts of FCR capacity for the upcoming week in an auction. Providers of FCR are compensated for capacity reservation based on the reserve power price, whereas delivered energy itself is not a matter of compensation.[2] Therefore, market players require appropriate forecasts of the week-ahead FCR power prices to be successful in the related auctions.

An individual supplier faces the tradeoff between the profit from selling FCR and the opportunity costs of the alternative use of flexible capacity, like bidding on the day ahead or the intraday market. Additionally, if the supplier decides to provide FCR, the technical unit has to be online for the entire week of provision. In the case of a power plant with minimum load requirements, the provider risks costs induced by negative contribution margins. Therefore, in order to prepare an adequate offer for the FCR tender and the other market segments, high-quality price forecasts are inevitable.

However, forecasting of FCR power prices has hardly been addressed in the forecasting literature (see Section 2). For this reason, we develop and introduce adequate forecasting models based on seasonal integrated autoregressive moving average (ARMA) models with exogenous regressors (SARIMAX) as explanatory variables and compare their results with methods from a second model family, the neural-network-based models. From the latter, we set up an experiment design to develop high-performing neural networks. The goal of this study is to find not only well-performing forecast methods but also their appropriate configuration in terms of hyperparameters and training strategies.

We find that both neural networks and SARIMAX models are capable of forecasting FCR prices reasonably well. For the neural networks, the simple network structures outperform the more sophisticated ones. The applied overfitting and ensembling techniques lead to significantly better forecast results and provide a solution to the problem of training data scarcity.

The main contributions and novelty of this paper are as follows:

1 Application and comparison of statistical and neural network models to price forecasting in reserve power markets that increasingly gain more importance in the energy transition era.
2 Description and discussion of training strategies for forecasting reserve power prices with neural networks on a scarce data basis.
3 Definition and discussion of appropriate target variable in the case of FCR prices in a market that is designed as a pay-as-bid auction.
4 Discussion on suitability and performance of simple and more sophisticated network structures for the mentioned market prices.

In this context, the paper is structured as follows. In Section 2, we review different approaches to forecast short-term electricity market prices in the literature. In Section 3, we deploy forecasting approaches considering autoregressive processes and exogenous drivers: precisely, a SARIMAX approach and artificial neural network models (ANN). Hereby, we consider feedforward units and set up an experiment design which deploys different model structures and training strategies. Finally, in Section 4, we apply the approaches to the stated forecasting problem and compare the performances. In Section 5, we conclude the findings and provide an outlook on future developments.

2 | RELATED LITERATURE

Among the first looking into the issue of reserve pricing and costs from a market perspective are Kirsch and Singh (1995). They provide an overview over the cost components of reserve power: opportunity costs of foregone sales, costs of uneconomic operation, potential startup and shutdown costs, costs resulting from frequent load changes and costs caused by efficiency losses. In addition, as applies for pricing electricity in the wholesale market, on the one hand the short-term marginal costs have to be considered. These are mainly determined by

[1]Note that France joined the procurement union in 2017 and subsequently provides more than a third of the required FCR. However, the market entry of France is considered in the model building, as the structural change may have introduced correlations and dynamics, which data from before 2017 do not contain.

[2]This is due to the fact that activation is hardly predictable and the delivered energy amount has an expected value of zero.

fuel and operation costs and can be increased due to partial load operation and decreased efficiency. On the other hand, the capital costs and other fixed costs need to be recovered by contribution margins generated in the market in the long term.

Weron (2014) finds that the actual modeling and forecasting of prices from balancing reserve and ancillary services markets has been comparatively rare in the literature. Exceptions include Olsson and Söder (2008), who model real-time balancing reserve power market prices in the Nordic market by using combined SARIMA and discrete Markov process models. They conclude that the developed model combination is suitable to use for the generation of real-time balancing power price scenarios. Klæboe, Eriksrud, and Fleten (2013) benchmark time-series-based forecasting models, and Dimoulkas, Amelin, and Hesamzadeh (2016) apply a hidden Markov model to forecast balancing reserve market prices for the Nordic market. They argue that activation of the balancing reserve occurs randomly and an activation-based price is therefore hardly predictable. Unfortunately, unlike the tenders considered in the present paper, the considered market design in the Nordic market is based on payments for reserve activation and not for the provision of reserve power.

Just and Weber (2008) consider an equilibrium model with two alternative competitive markets: the secondary balancing reserve power and an hourly electricity spot market. They valuate the provision of balancing reserve by quantifying the opportunity to spot market sales and deduce a development of capacity prices for secondary balancing reserve power for the German case. However, they do not apply the equilibrium model to forecast prices and do not include FCR in their investigations.

Finally, Wang, Zareipour, and Rosehart (2014) investigate the application of established stochastic approaches for modeling the behavior of operating reserve and regulation prices in the North American electricity markets, which, like the Nordics, are based on activation rather than provision of balancing reserve power. The investigated models are descriptive and not designed for generating short-term forecasts. The authors point out that reserve and regulation prices are characterized by higher volatility, lower mean, more frequent price spikes, and a more skewed distribution compared to electric energy prices. Thus modeling reserve power prices is potentially more challenging.

In contrast to forecasting reserve market prices, forecasting of electricity spot market prices is a field that has been pervasively studied (Weron, 2014). For example, Kiesel and Paraschiv (2017) and Bublitz, Keles, and Fichtner (2017) mention fundamental price drivers such as load, fuel prices, unavailable generation capacity, and renewable energies' feed-in as suitable exogenous regressors to forecast electricity prices.

ANN forecasting of hourly day-ahead electricity prices and a comparison to econometric benchmarks was first applied by Catalão, Mariano, Mendes, and Ferreira (2007), who find a good forecasting performance of ANN on the Spanish and the Californian market. Lago, Ridder, and Schutter (2018) study the Belgian day-ahead electricity market and consider a large set of possible forecasting models, concluding a significant dominance of machine learning over the statistical models in terms of forecasting accuracy. Ugurlu, Oksuz, and Tas (2018) and Oksuz and Ugurlu (2019) forecast the Turkish day-ahead and intraday market electricity prices with different neural networks configurations, including feedforward, gated recurrent unit (GRU) and long short-term-memory (LSTM) model designs. The authors conclude a significant dominance of GRU model designs and state an improvement with increasingly sophisticated network structures. Giovanelli, Sierla, Ichise, and Vyatkin (2018) forecast the hourly day-ahead balancing prices of the Finnish market and compare neural networks in various parameter configurations with support vector regression and autoregressive integrated moving average (ARIMA) models. They find that the amount of training data is a key impact on the forecasting performance of the models, whereas different training strategies, algorithms, and activation functions performed similarly well.

The methodological approach of comparing models originating from econometrics with machine learning models has been applied to several scopes in the literature. Chatfield (1996) and Adya and Collopy (1998) provide a theoretical foundation for the need to consider both econometric models and machine learning approaches such as neural networks in forecasting. They conclude that the model setup requires a careful choice of external regressors with regard to out-of-sample-fit in order to respect model uncertainty (Chatfield, 1996) and that well-designed ANN models have the potential to outperform econometric approaches in forecasting applications (Adya & Collopy, 1998). Early studies deploying both econometric and ANN models include applications in forecasting electricity demand (Liu et al., 1991), consumer expenditure (Church & Curram, 1996), retail sales (Alon, Qi, & Sadowski, 2001), foreign exchange rates (Qi & Zhang, 2001; Yao & Tan, 2000), gross domestic product (GDP) growth (Tkacz, 2001), stock returns (Olson & Mossman, 2003; Qi & Zhang, 2001), and inflation rates (Binner, Bissoondeeal, Elger, Gazely, & Mullineux, 2005). The studies confirm the conclusion regarding the forecasting potential of well-designed

neural networks drawn by Adya and Collopy and suggest the adaption of the study design to FCR price forecasting.

However, for all mentioned studies the data basis for training the model is comprehensive. In particular, the studies on electricity prices rest on hourly data of several years and the neural networks thus have plenty of observations to learn from. Further, spot market prices are typically well explainable by fundamental factors (see, e.g., Bublitz et al., 2017; Kiesel & Paraschiv, 2017; Weron, 2014). Conversely, a challenge in forecasting balancing reserve market prices lies in the fact that they are hardly explainable by fundamental drivers (Kraft, Keles, & Fichtner, 2018; Ocker, Ehrhart, & Belica, 2018). However, Ocker and Ehrhart (2017) find evidence for collusion among market participants and serial correlation in the auction results of the secondary reserve market. Another key challenge in this paper is based on a relatively sparse database, consisting of weekly data from the years 2017 and 2018. To cope with the data scarcity, we deploy ensembling and overfitting strategies (see Section 3) that, to the best knowledge of the authors, have not been deployed in electricity price forecasting before.

We are well aware that commercial providers offer forecasts for the considered FCR market. Unfortunately, however, these commercial providers publish neither their methodologies in detail nor historic forecast time series as a benchmark. In the next section, we will therefore follow Weron (2014), who classifies short-term price forecasting models into time series analysis approaches and artificial intelligence or machine learning approaches. We will set up and deploy forecasting models for the FCR price based on both time series analysis (SARIMAX) and ANN.

3 | METHODOLOGY

The literature review in the previous section displayed a lack of scientific publications in the field of FCR price forecasting and suggested the application of, on the one hand, approaches coming from time series analysis, and, on the other hand, approaches coming from machine learning. To obtain a benchmark that is neither time series based nor machine learning based, a naive forecast[3] is taken as a benchmark. Preliminary analyses showed that for FCR prices the naive forecast outperforms linear regression and can well compete with a SARIMA approach (Kraft, Rominger, Mohiuddin, & Keles, 2019). In Section 3.1, owing to the pay-as-bid auction design, first the dependent variable is defined and its time series is analyzed briefly. In Section 3.2, the exogenous variables required for the forecasting approaches are introduced and their preprocessing is explained. Sections 3.3 finally presents the setup and training of the SARIMAX and ANN models.

3.1 | Definition of dependent variable and time series analysis

As FCR tenders are pay-as-bid auctions, there is no uniform settlement price but each market participant receives its price bid as remuneration. Prior to setting up a highly sophisticated forecasting model, it is necessary to define a suitable dependent variable. Analysis of the FCR market results from 2014 to 2018:Q3 (Figure 1) shows the range of accepted bids as well as the capacity-weighted average price in each auction. From the relatively low gap between the capacity-weighted average price and the respective marginal price (except for single spikes), we conclude that the capacity-weighted average is a suitable target variable for the forecast. The main errors induced by using the capacity-weighted average instead of the maximum price, for example, arise in periods with price spikes. However, considering that a risk-neutral trader would not speculate on the height of price spikes, the capacity-weighted average remains the favorable forecast target.

The time series contains seasonality, mainly induced by a strong price increase over the Christmas holidays and a moderate price increase in early summer of each year. In general, the time series shows a decreasing trend. To check the time series for stationarity, it was tested with Kwiatkowski–Phillips–Schmidt–Shin (KPSS) unit root tests (Kwiatkowski, Phillips, Schmidt, & Shin, 1992). The nondifferenced time series rejects the stationarity null hypothesis at 1% significance; the series of first differences (shown in Figure 2) does not reject the stationarity null hypothesis. The econometric models will therefore be estimated with a SARIMAX approach with

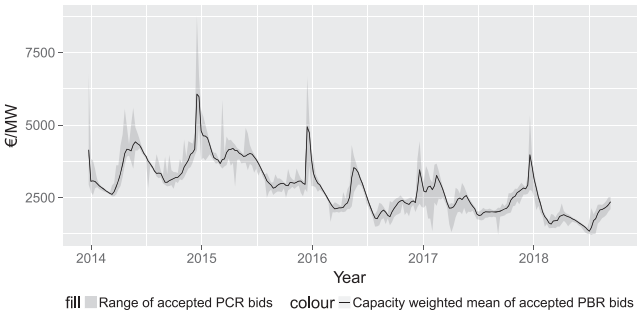

FIGURE 1 FCR price development from 2014 to 2018:Q3 (own illustration based on data from regelleistung.net, 2019)

[3]The naive forecast equals the expectation of having the same price as in the last auction.

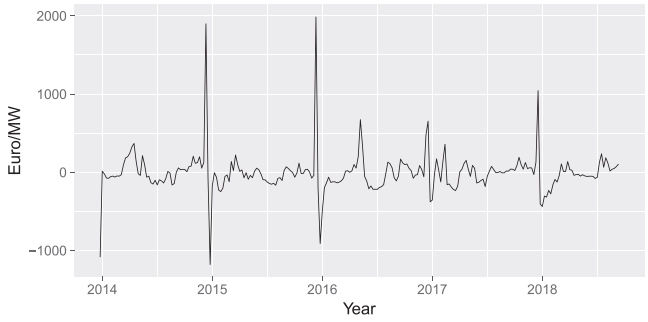

FIGURE 2 First differences of capacity-weighted average of accepted FCR bids from 2014 to 2018:Q3 (own illustration based on data from regelleistung.net, 2019)

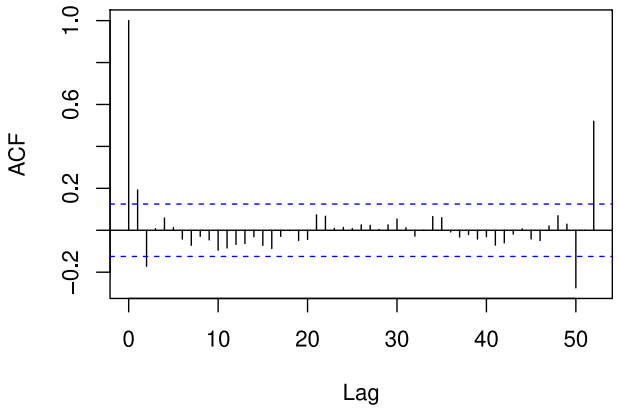

FIGURE 3 Autocorrelation function (ACF) of differenced time series (own illustration based on data from regelleistung.net, 2019) [Colour figure can be viewed at wileyonlinelibrary.com]

the undifferentiated time series of capacity-weighted averages as dependent variable y. The SARIMAX approach allows us to endogenously model the first differences Δy of the time series in order to derive forecasts regarding the forecast target. The autocorrelation function (ACF) of the differenced time series indicates a significant correlation with lag 1, lag 2, lag 50, and lag 52 (see Figure 3). Thus, for model training and prediction, Δy_{t-1}, Δy_{t-2}, Δy_{t-50}, and Δy_{t-52} are supplied as the respective lags.

For the ANN models, the dependent variable is defined as the first difference Δy of the capacity-weighted price time series, corresponding to the difference between the price of the current and the price of the previous auction. In order to return to the desired FCR price prediction, the predicted difference is added to the FCR price of the previous auction. This procedure complies intuitively with the SARIMAX model, which likewise intends to estimate the first differences instead of the actual forecast target, and is therefore considered a suitable comparative approach.

Table 1 summarizes the statistical properties of mean, median, standard deviation, skewness, and kurtosis for the times series of the differences of FCR prices in the period of investigation 2017[4] to 2018:Q3 with a total number of 88 observations.

3.2 | Identification and pre-processing of exogenous variables

As there is no explicit literature on exogenous regressors with regard to balancing reserve prices, several regressors that are commonly used in models for other electricity prices (see, e.g., Bublitz et al., 2017; Kiesel & Paraschiv, 2017) are considered as exogenous regressors in this study. Representing, among others, opportunity costs for reserve provision and a scarcity in the market, the following possible predictors are identified:

- price range and skewness of FCR bids in previous auction (regelleistung.net, 2019);
- average electricity price of week-ahead future German–Austrian (DE-AT)[5] and French (FR) market area (EEX, 2019);
- average day-ahead electricity spot market price in DE-AT and FR (EEX, 2019);
- average load forecast and realized load for DE-AT and FR (ENTSO-E, 2019);
- number of German public holidays in a week (ENTSO-E, 2019);
- planned unavailable capacity in DE-AT and FR (ENTSO-E, 2019).

Note that exogenous factors like wind and photovoltaic power feed-in are not considered, as the auction for FCR procurement takes place 1 week ahead and the volatile renewable feed-in is hardly predictable at these timescales. However, the future price includes the effect of the expected wind and photovoltaic power feed-in in the respective week due to the merit-order effect. We thus implicitly consider for volatile renewable energy sources to some extent.

For the selection of predictors from the list above, the *corrected* Akaike information criterion (AIC; Hyndman &

[4]As France joined the joint auction at the start of 2017, data from before that date may not include all interdependencies and lead to a wrong model fitting.

[5]As the DE-AT future product was split up into DE and AT future products, the volume-weighted average of DE-AT and DE futures is taken for 2018.

TABLE 1 Descriptive statistics of the differences of FCR prices

Variable	*n*	Mean	Median	SD	Skewness	Kurtosis
Differences, FCR price	88	−8.13	9.62	178.58	1.91	15.38

Athanasopoulos, 2013) of a linear regression model[6] applied to 2017 data is used. Other popular information criteria for model selection contain the regular AIC and the Bayesian information criterion (BIC). By penalizing the number of parameters, the corrected AIC accounts for and adjusts the tendency of the AIC to prefer models with too many parameters when sample sizes are relatively low. Due to the relatively low sample size, the AIC was not considered in predictor selection. By penalizing the number of parameters, the corrected AIC accounts for and adjusts the tendency of the AIC to prefer models with too many parameters when sample sizes are relatively low. Among all predictor combinations, the set of exogenous predictors containing the *FCR price range*, the *future price DE-AT*, the *future price FR*, the *load in DE-AT*, the *load in FR* and the *planned unavailable capacity in DE* achieved the lowest corrected AIC, corresponding to the best fit in the linear regression on the 2017 data. A selection based on the BIC leads to a similar parameter set as the ranks of the models sorted by BIC are comparable to the ranks sorted by the corrected AIC. For example, the best model in terms of BIC chooses the *load forecast in FR* instead of the *realized load in FR* and drops the *future price of FR*. For the scope of the paper to configure and compare the SARIMAX and ANN forecasts, we consider the choice of regressors according to the corrected AIC to be suitable. As French nuclear power plants contribute a significant share to the FCR provision, the *planned unavailable capacity in FR* is added to the predictor set chosen by the corrected AIC. Although the chosen predictor set x may not be the best for all models, all forecasting approaches are deployed in the following with the same selected set for reasons of consistency and comparability.[7]

[6]Linear regression models the differences of the FCR price time series (dependent variable = Δy) with the different sets of exogenous variables. The regression was chosen over a simple correlation analysis as the latter might not respect interdependencies between the independent variables. In particular, load and electricity prices are highly correlated and should thus not be handled independently.

[7]The predictor set containing the *planned unavailable capacity* in France instead of the *planned unavailable capacity in Germany* was the eighth best (of 16,383) behind variations of the highly correlated *load* and *load forecast in Germany* and *in France*. The corrected AIC penalizes adding a predictor to the set; thus the predictor set finally used was not among the favorites of corrected AIC. Nevertheless, as mentioned above, we consider the *unavailable capacity in France* a relevant predictor variable and included it in the investigation.

The preprocessing consists of a validity check of the raw data, the calculation of descriptives to be used in the modeling (e.g. weighted average, range or skewness), and finally a normalization. Normalization has been discussed at many points in the context of time series forecasting and neural networks (see, e.g., Kaastra & Boyd, 1996; Keles, Scelle, Paraschiv, & Fichtner, 2016). For ANN, it is particularly important to choose the normalization range according to the intended activation function of the neurons. As having a common value range of all target and predictor variables leads to a more stable functioning of the related fitting algorithms and does not change the results, we normalize the data between zero and one by subtracting the minimum value and dividing by the range of values.

3.3 | Setup and training of models

For training and forecasting with the SARIMAX and ANN models, a cross-validation approach called *rolling one-step forecast with model reestimation* is set up (see, e.g., Arlot & Celisse, 2010). In this approach, models are fitted with training data in order to predict the value of the single next step. In the reference training strategy the training data set is extended by one step for each forecast step, which is also referred to as an expanding window. In our case, the initial training data set consists of the 52 observations from 2017. As can be seen in Figure 4, the training data set for week 1 of 2018 consists of all 2017 data, the training data set for week 2 of 2018 consists of the 2017 data plus week 1 of 2018, and so on. In this way, the best information available to the trader at the forecasting time is used in the forecast. As a consequence, there is no single model but as many models as forecasting steps for each approach presented in the following paragraphs.

As the analysis of the price time series in Section 3.1 revealed different price characteristics over time, in addition to the expanding window, a rolling window of size 10 is considered in the experimental design for training the ANN. The rationale behind having a rolling window is to make the networks more adaptive to changing dependencies over time and to focus on the recent observations, not distorting the network learning from nonrelevant information from the past. However, the rolling window obviously bears the risk of further enhancing the data scarcity problem and leading to worse prediction

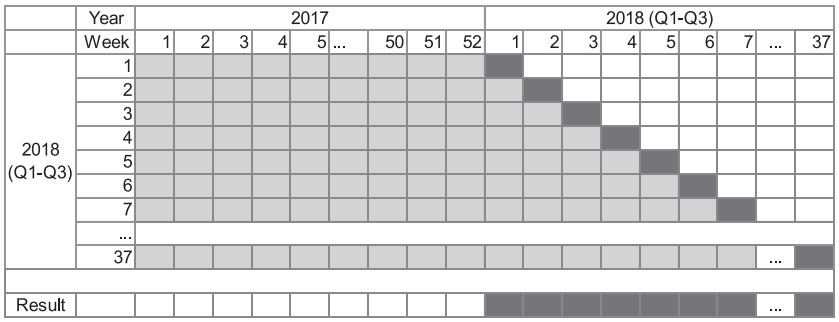

FIGURE 4 Visualization of the rolling one-step forecast with model reestimation and expanding window (own illustration). Time steps in light gray mark the training data that are used for forecasting the dependent variable in the time steps (in dark gray)

results as well as less robust models. In this way, the strength of more sophisticated network structures cannot be exploited in the same way that is possible with a larger training data set.

3.3.1 | SARIMAX model

The setup of a SARIMA(p,d,q) $(P,D,Q)_m$ model consists of defining the optimal values for the hyperparameters (Brockwell & Davis, 2016):

- p: trend autoregression order;
- d: trend difference order;
- q: trend moving average order;
- P: seasonal autoregression order;
- D: seasonal difference order;
- Q: seasonal moving average order;
- m: time steps for a single seasonal period.

In addition, to apply a SARIMAX model we include the exogenous predictor variables x presented in Section 3.2 with the same order as the trend autoregression order.[8] As mentioned in Section 3.1, the KPSS test suggests a difference order $d = 1$. The ACF of the differenced time series indicates a significant correlation with lag 1, lag 2, lag 50 and lag 52 (see Figure 3). Therefore, the SARIMAX model is set up with the trend autoregression order $p = 2$ to consider lags 1 and 2; lags 50 and 52 are considered by the seasonal hyperparameters. For the remaining hyerparameters, a parameter grid search is performed to fit the optimal model of SARIMAX class to the training data by deploying a variation of the Hyndman–Khandakar algorithm. It deploys and combines iterative unit root tests, minimization of the AICc and maximum-likelihood estimation to obtain the optimal model order. The training of the models is a side effect of the used algorithm and based on maximum-likelihood estimation. For further details, see Hyndman and Khandakar (2008).

[8]With $p = 2$ leading to x_{t-1} and x_{t-2} as model input.

3.3.2 | Artificial neural networks

Whereas the econometric approach fits the model to the training data by assuming linear relationships between inputs and outputs, ANN are capable of training nonlinear relationships of the input data to explain the variations in the dependent variable. In the following paragraphs the configuration and training of the ANN are presented, which comprises the description of the important hyperparameters—*neurons per hidden layer*, *number of hidden layers*, which *training data* to select, and which *training hyperparameters* to apply—and their variations in the experiment design. In a first step, the ANN are set up by defining the network structure and units. In a second step, training strategies are defined to specify the way the training data are processed through the networks. However, there are infinite possible combinations of network configurations as well as training strategies.

In the following, we present a reference model configuration and an experimental design with variations in the number of hidden layers and number of neurons per layer. As the application of forecasting FCR prices offers – compared to other applications of neural networks such as picture classification or language processing – few training data, in this paper multilayer perceptron feedforward models are deployed. In the working process, further advanced model developments such as recurrent network structures with GRU as neurons were also assessed but yielded no improvement. Therefore, they are not presented in the following. For the sake of completeness, the hyperparameters and training configurations and results of the GRU forecasts are provided in Tables A and B in the Appendix.

As a reference, a feedforward model with one hidden layer and 10 neurons using a rectifier activation function, often referred to as rectified linear units (ReLU),[9] is

[9]Rectifier activation functions have become very well known in deep learning recently, outperforming the more known logistic and hyperbolic tangent activation functions (see, e.g., Glorot, Bordes, & Bengio, 2011; LeCun, Bengio, & Hinton, 2015; Ramachandran, Barret, & Quoc, 2017).

configured. The choice of the number of *neurons per hidden layer* is important in the setup of a network. Generally, there is no optimal model configuration algorithm, but there are many rules-of-thumb. One of them suggests a number below the half of input nodes and approximately two thirds of the sum of input and output nodes. Although this is only a rule-of-thumb and the choice is problem specific, we conclude 10 to be a reasonable number of neurons per hidden layer for the reference model. To gain insight into the sensitivity of the number of neurons per layer, in the experimental design the design variable is varied with levels 10 and 20.

The second hyperparameter choice in the model configuration is the *number of hidden layers*. For the reference model, one hidden layer is chosen. Originating from the structure, deeper networks are more adaptive to the training data and thus able to learn more complex relationships. A drawback of deep networks, especially those that are trained with relatively few data, is the risk of overfitting. To investigate the dependency of the prediction on the amount of layers, a second configuration with two hidden layers is deployed.

As important as the configuration of the model structure is the definition of the training strategy. In this paper, the term *training strategy* comprises the selection of *training data* to be available and the way these are processed in the training process, defined by the *training hyperparameters*. To provide the networks with the same input data as the SARIMAX approach, for each prediction a lookback of two steps into the data is implemented, containing the values of the dependent variable and their lags of 50 and 52 as well as seven exogenous variables, leading to 22 input nodes.

The last stage before training and evaluating the networks is the definition of the *training hyperparameters*. As applies for the network configuration, the training is subject to the tradeoff between utilizing all information that is contained in the training data and overfitting the model to the training data. Training the models requires the setting of the hyperparameters *number of batches*, *number of epochs*, and *iterations per epoch*, defining the way the training data set is split into training batches and how the training in terms of weight optimization is executed. As our models face relatively small data sets, batching the training data is not necessary (*number of batches = 1*).

For the other hyperparameters, *number of epochs = 30* (representing the number of training sequences), and *iterations per epoch = 20* (representing the number of iterations optimizing the tensor weights per sequence) lead to favorable training results for the reference model configuration with the expanding window training data outlined above. As shown in Figure 5, the choice of

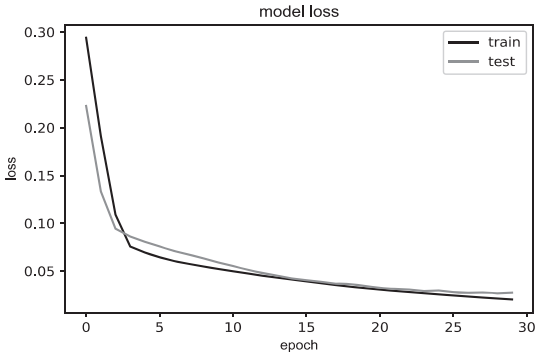

FIGURE 5 Exemplary training history of the reference feedforward network with training hyperparameter set "fit" (number of batches = 1, number of epochs = 30, iterations per epoch = 20). The black line depicts the loss of model training, and the gray line depicts the loss of a random 10% validation split extracted from the training data

hyperparameters yields a desirable training fit and avoids overfitting. Hereby, due to the rolling one-step forecast with model reestimation setup, we consider it suitable to take a 10% validation split randomly selected from the training data and do not apply a hold-out-sample validation. The validation split is only conducted for the hyperparameter selection. After the hyperparameters are selected, owing to data scarcity the validation split is dropped for model training. The training of the ANN is thus conducted with the entire training data set to account for all relevant information.

Since the model training starts with random weights and is therefore indeterministic, an ensemble of networks is deployed for each configuration. Ensembling is a common technique similarly proposed by Hyndman and Athanasopoulos (2013). For applications with rich databases (see Section 2 for examples) for model training and validation, ensembling is not very important as model training mostly converges to a single model. However, in our case with a scarce training data basis, ensembling allows us to obtain robust forecasts from numerous indeterministic models. For the reference training hyperparameters, we run the fitting process 50 times to obtain 50 independent ANN of each model structure for each forecasting step. The prediction values of these are then averaged to obtain a single representative prediction value for the respective forecast step. As a measure of robustness, the standard deviation of the different forecasts within the ensemble is reported in Section 4. Although different hyperparameters could yield better training results, they also bear the risk of overfitting the data.

An alternative training strategy consists of intentionally overfitting the training data to some extent and compensating the overfit by increasing the *ensemble size*. To

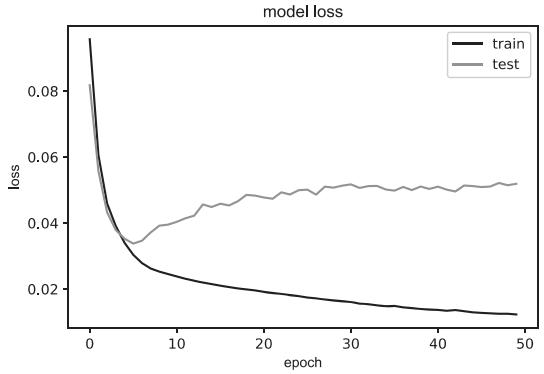

FIGURE 6 Exemplary training history of the reference feedforward network with training hyperparameter set "overfit" (number of batches = 1, number of epochs = 50, iterations per epoch = 30). The black line depicts the loss of model training, and the gray line depicts the loss of a random 10% validation split extracted from the training data

examine the performance of this strategy compared to the reference "*fit*" hyperparameter set, a second set of *training hyperparameters* "*overfit*" is implemented. Increasing the *number of epochs* to 50 and the *iterations per epoch* to 30 leads to a slight overfit for the reference model configuration. As the exemplary training history in Figure 6 shows, the performance of the model on the validation data becomes worse as the model fit increases with advancing training. However, equipped with an *ensemble size* of 100, these intentionally overfit models might, on average, perform better than the fit models as the overfitting residuals balance each other out.

In the end, the presented experiment design with four factors and two levels each leads to 16 different model configurations. Table 2 summarizes the factors and their levels in the network configuration and training process. In the following, the abbreviation for a combination of a network configuration and a training strategy is built by combing the entries of Table 2—for example, "FF1_10_E_F" for the reference with one hidden layer, 10 neurons per hidden layer, an expanding training window, and the training hyperparameters "fit."

The training and evaluation of the ANN models are implemented in keras,[10] a common machine learning library available for Python and R. On a machine with a 2.50 GHz 64-bit processor (central processing unit, CPU) and 16 GB RAM, depending on the model configuration and training strategy, the training and evaluation of one setup takes between 2 and 6 hours for the 37 forecasting steps. However, the training time could be significantly reduced by the use of parallelization and a graphics processing unit (GPU) for the computations.

[10]For more information on keras see https://keras.io/.

4 | RESULTS

The results consist of the out-of-sample performance of the presented model framework. The forecasted time series for a selection of approaches in comparison to the real time series of FCR prices of the testing period (01/2018–09/2018) are presented in Figure 7. The selection consists of the naive forecast, the SARIMAX approach, as well as ANN approaches "FF1_10_E_F," "FF1_20_E_F," "FF1_10_E_Ov," "FF1_20_E_F," "FF2_10_E_F," and "FF2_10_E_Ov." Due to conciseness, the plots for the complete set of examined configurations of the ANN experiment design have been moved to Figure A in the Appendix.

A first graphical comparison of the developed forecasts (Figure 7) indicates that both the econometric and ANN approaches are able to forecast the level of the FCR price quite well. For the SARIMAX approach, both the point forecast and the 95% confidence intervals are provided. The latter indicate the robustness of the estimated model for each forecast step. It can be observed that for all time steps the prediction target (dashed line) lies within the confidence interval, complying with a desirable robustness. As there is no mathematical equivalent for confidence intervals in the ANN approaches, the robustness of the models is determined with the standard deviation of the point forecasts obtained from ensembling for each forecast step (provided by Table 3). An artefact confidence interval for the ANN approaches could be constructed from the residuals' distribution of sufficiently large ensembles (e.g. 1,000 networks instead of 50). The residuals would then represent an empirical distribution, whose 2.5% and 97.5% quantiles could be interpreted as the confidence interval. However, computational limitations do not allow us to generate ensembles of size 1,000 for each forecasting step and all model designs. We cannot build a reliable distribution for the residuals based on an ensemble of size 50. Without a reliable distribution, no confidence interval in a mathematical sense can be derived and the construction of confidence intervals for ANN is excluded.

The good fit also counts for the naive approach, so that the benefit of the more sophisticated models does not become clear at the first inspection of the results. A second view reveals that the deviation between the forecasted values and the real test data is especially smaller when the overall price level and in particular the price variations decrease. For the high price levels (first parts of the price curve), it is observable that the ANN approaches perform better and almost approach the real price curve. These differences become more visible if the residuals—that is, the single forecast errors—are directly analyzed. Figure 8 shows that, compared to the naive and

TABLE 2 Experimental design for neural networks

Factor	Level reference	Level variation
Network configuration		
Number of hidden layers	FF1 (1 hidden layer)	FF2 (2 hidden layers)
Number of neurons per hidden layer	10	20
Training strategy		
Training data	E (expanding window)	R (rolling window)
Training hyperparameters	F (fit): number of batches = 1, number epochs = 30, iterations per epoch = 20, ensemble size = 50	Ov (overfit): number of batches = 1, number epochs = 50, iterations per epoch = 30, ensemble size = 100

Note. A design consists of a combination of the hyperparameters' number of hidden layers, number of neurons per hidden layer and the training strategy defined by the training data and the training hyperparameters

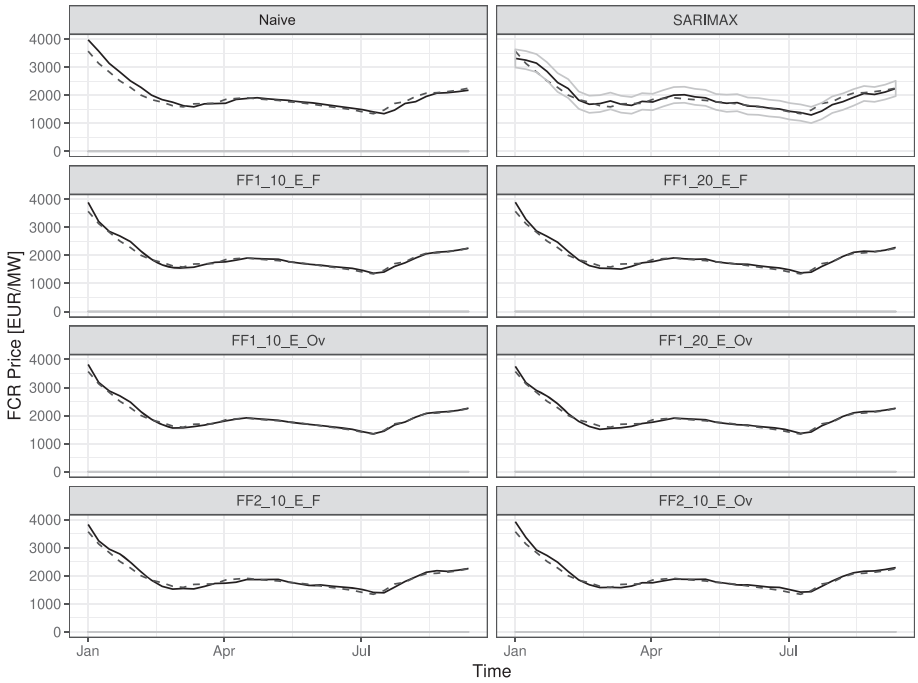

FIGURE 7 Selection of FCR price forecasts in test period 2018:Q1–Q3 (original FCR price data from regelleistung.net, 2019). Solid lines are the forecasted FCR prices, and dashed lines represent the realized FCR price. All forecasting approaches show a relatively good fit

the SARIMAX approach, the errors of the ANN approaches are particularly lower for the first part of the test period (until April), when real FCR prices have a strong decline and are exposed to more fluctuations. The residuals indicate a serial correlation that was also reported in more detail in the preliminary works of Kraft et al. (2019). Generally, the residuals of well-fitted SARIMAX models should be independent and identically distributed. In the rolling one-step forecast with model reestimation setup deployed in this paper, each forecasting step reestimates the model, which leads to distinct SARIMAX models for each forecasting step.

Further investigations address the structure of the SARIMAX residuals to check for conditional heteroskedasticity. Figure 8 therefore exemplarily provides the residuals of the SARIMAX model estimated for the last forecasting step. It can be observed that the residuals are not perfectly homoskedastic. Although no substantial autocorrelation is observable, the volatility of the time series appears to be heterogeneous over time. The residuals in 2017 are larger compared to those in 2018 and, in particular, the year change from 2017 to 2018 produces two data points with a larger volatility compared to the rest of the time series. To address the suspected heteroskedasticity of the SARIMAX residuals, a SARIMAX–generalized autoregressive conditional heteroskedasticity (GARCH) approach was tested. However, the limited data basis

TABLE 3 Root mean square error (RMSE), mean absolute percentage error (MAPE), directional accuracy (DAC), and mean standard deviation (σ) of the model forecasts

Design	RMSE	MAPE	DAC	σ
Naive	158.16	5.24%	91.70%	n/a
SARIMAX	136.82	5.18%	75.00%	140.03
FF1_10_E_F	**86.38**	**2.78%**	**100.00%**	**127.81**
FF1_20_E_F	94.13	3.27%	91.70%	125.75
FF1_10_E_Ov	**72.16**	**1.97%**	**97.20%**	**108.05**
FF1_20_E_Ov	**72.71**	**2.89%**	**97.20%**	**120.78**
FF1_10_R_F	185.71	6.32%	66.70%	183.35
FF1_20_R_F	190.94	6.52%	75.00%	175.02
FF1_10_R_Ov	194.97	6.43%	72.20%	178.68
FF1_20_R_Ov	194.05	6.80%	72.20%	167.71
FF2_10_E_F	101.42	3.94%	80.60%	147.43
FF2_20_E_F	119.77	4.75%	77.80%	134.81
FF2_10_E_Ov	104.07	3.45%	86.10%	147.30
FF2_20_E_Ov	114.96	4.79%	80.60%	128.12
FF2_10_R_F	181.49	6.22%	69.40%	158.77
FF2_20_R_F	189.05	6.23%	72.20%	131.80
FF2_10_R_Ov	192.37	5.74%	80.60%	142.14
FF2_20_R_Ov	184.20	6.02%	77.80%	120.63

Note. For ANN the reported σ is calculated as the empirical standard deviation of residuals, whereas for SARIMAX σ is the mean theoretical σ of the 37 forecast models. The simplest design FF1_10_E_F reaches 100% of DAC, but is dominated by FF1_10_E_Ov and FF1_20_E_Ov in terms of RMSE. The best design by RMSE and MAPE is FF1_10_E_Ov. These three designs are indicated in bold font. The more sophisticated designs and the designs with a rolling training window have a similar performance to the SARIMAX and naive forecast. All designs involving ensembling show a moderate standard deviation, indicating robust model training and the need for ensembling

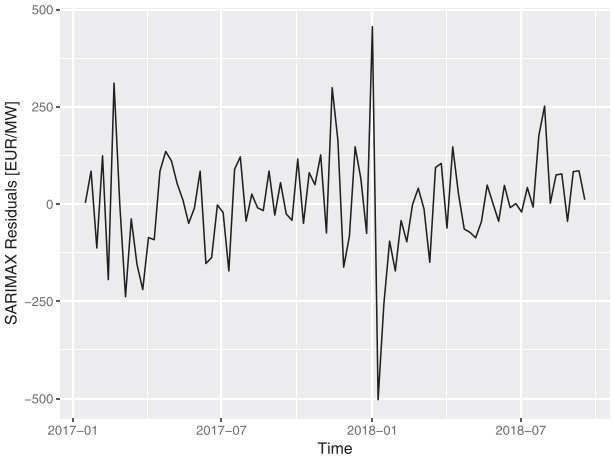

FIGURE 8 Exemplary residuals time series of the SARIMAX model for the last forecasting step. It can be observed that the residuals do not contain substantial autocorrelation and that the volatility of the residuals is increased at the year change from 2017 to 2018

(residuals contain only 52–88 observations for the different forecasting steps) impedes the deployment of GARCH models, as the estimation does not converge. Related literature suggests sample sizes of at least 500 (respectively 700) are required to obtain good results for GARCH volatility estimation (Hwang & Valls Pereira, 2006; Ng & Lam, 2006). Unfortunately, in our case the data basis is too scarce to apply GARCH and we are restricted to the chosen SARIMAX approach as econometric comparative to the ANN.

The naive approach is performing similarly well (respectively even better) in periods when prices remain more or less constant over time. However, this is quite obvious, as this approach applies the last week's real value to the current week's forecast. In periods with hardly any changes, the approach will therefore produce desirable results.

However, we are more interested in approaches that can also capture periods when prices undergo price changes, as future FCR prices might change much more frequently and in a more pronounced way. The market is more and more opened for new players and technologies, such as battery storage, that will bring much more dynamics into the market. In this respect, the ANN approaches are able to capture price dynamics, which obviously cannot be covered by the naive approach. Moreover, in the case of FCR prices, ANN approaches cover the dynamics in volatile periods significantly better than the applied SARIMAX approach.

Interestingly, Figure 8 demonstrates that SARIMAX errors are more frequently fluctuating around 0 EUR/MW, while those of the ANN forecasts remain in the positive or negative scale longer. As Figure B in the Appendix indicates, the fluctuations around 0 EUR/MW also apply to the ANN configurations with a rolling training window, and for some forecasting steps a rolling window training strategy can yield a better forecast than an expanding window approach. This indicates that, by including seasonal factors or limiting the training data to a rolling window, the performance of the ANN approaches can be improved for some time periods. However, determining such factors based on the short period, for which weekly FCR prices are available, leads to overfitting for other periods, and therefore does not improve the overall forecasting performance itself. The concerns regarding the impeded exploitation of strengths of the ANN approach that are discussed in Section 3.3 prove

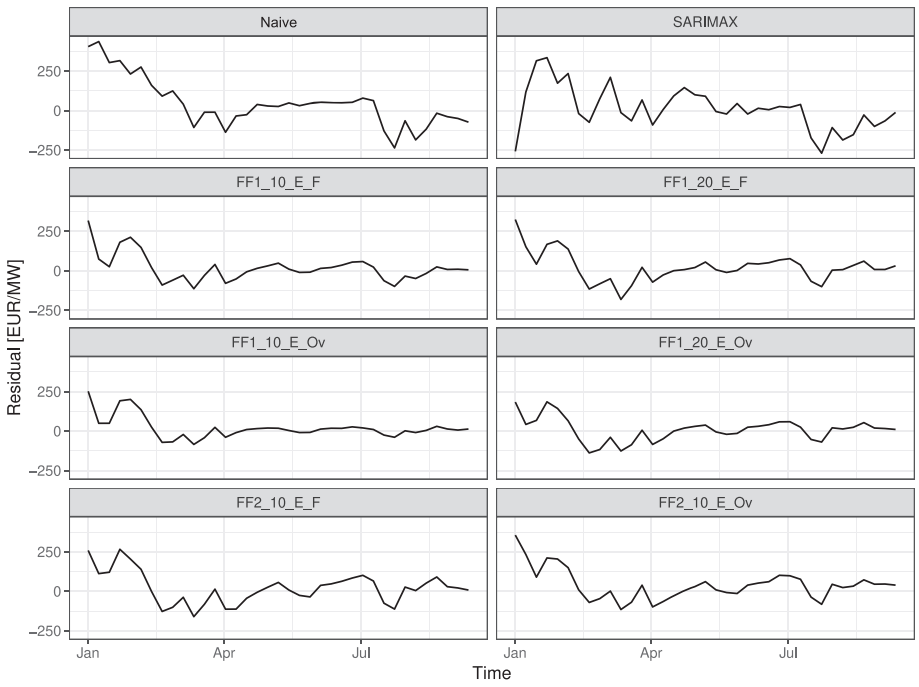

FIGURE 9 Residuals of FCR price forecasts in test period 2018:Q1–Q3 (original FCR price data from regelleistung.net, 2019). Due to the larger ensemble size, the ANN approaches with overfit (Ov) produce smoother residuals compared to those with fit (F)

right. As the rolling window training strategies cannot exploit the entirety of training data provided to the expanding window training, the overall forecasting performance decreases. Training with an expanding window is therefore preferable to the rolling window approach. In particular, if the network structures become more sophisticated, the model requires as many training data points as possible to be performant.

The residuals of the ANN approaches are distributed relatively symmetrical around zero, as can be seen in the illustration of error histograms in Figure C in the Appendix. However, to derive further insights regarding the distribution of the residuals, the number of forecasting steps is too small.

Whereas Figures 7 and 9 enable a qualitative discussion, Table 3 presents the quantitative performance and robustness measures. The performance measures again indicate that having a mean absolute prediction error (MAPE) below 7% all proposed models perform reasonably well. With regard to the root mean square errors (RMSE) and the MAPE, the feedforward ANN with an expanding training window all outperform the naive forecast and the SARIMAX models. The directional accuracy (DAC) confirms these observations. Whereas the best model in terms of RMSE and MAPE fails to predict the direction of change once (97.2% accuracy), the model design FF1_10_E_F reaches 100% accuracy in the considered forecasting steps.

Surprisingly, adding a second layer to the networks does not improve the forecasting. The dominating designs for the prediction task are the ANN with one hidden layer and an expanding training window. With an RMSE of 72.16 and a MAPE of 1.97%, the configuration with 10 neurons per layer and the "overfit" training (FF1_10_E_Ov) yields the best results. Increasing the number of neurons to 20 (FF1_20_E_Ov) or changing the training strategy to "fit" (FF1_10_E_F) leads to slightly worse results. Interestingly, the FF1_20_E_Ov design dominates the FF1_10_E_F design in terms of RMSE but is outperformed in terms of MAPE, meaning the residuals are on average larger but have smaller large residuals, which is penalized more strongly in the RMSE measure.

The design variable *neurons per layer* reveals an interesting, yet intuitive, pattern. Apart from the best-performing design with one hidden layer, expanding window and overfit training, the forecasts of the networks with 20 neurons per layer are more robust than the comparable networks with only 10 neurons. However, except for two cases with rolling window, the performance in terms of RMSE and MAPE is better for the configurations with only 10 neurons per hidden layer. The increased number of neurons leads to more convergence in the weight optimization and thus to more stable results, yet the convergence may be prone to overtraining of the relationships in the training data compared to the simpler configurations with only 10 neurons.

The last variable in the experimental design are the training hyperparameters, for which the two sets "fit" and "overfit" are distinguished. For the simple networks (one hidden layer, 10 or 20 neurons), the strategy to overfit and build a larger ensemble yields a massive

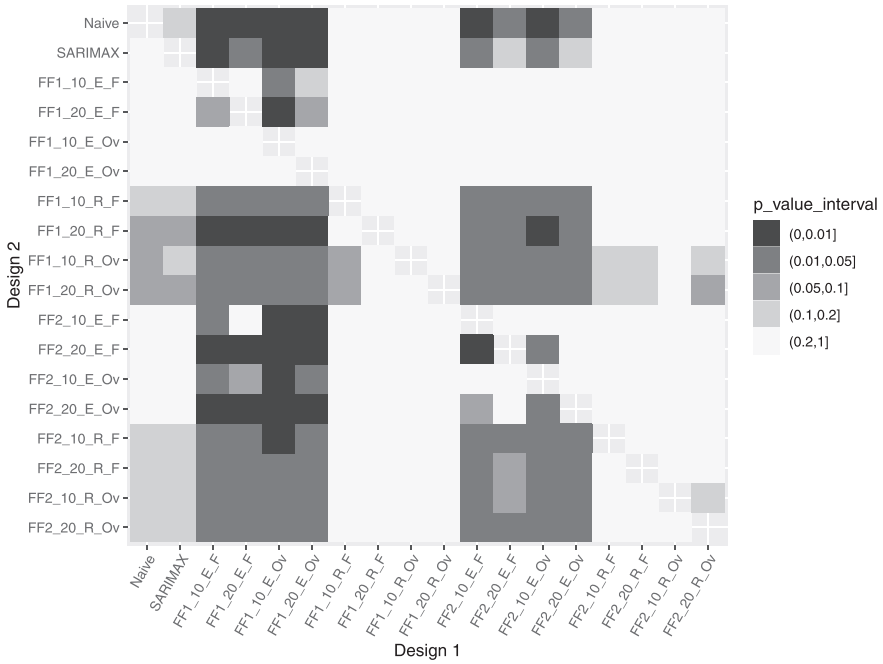

FIGURE 10 Results of Diebold–Mariano testing whether the forecasts obtained from design 1 are significantly better than the forecast obtained from design 2. It can be observed that the designs with expanding training window dominate those with rolling training window. Only the best three models (FF_10_E_F, FF_10_E_Ov, FF_20_E_Ov) are better than the SARIMAX approach at the 1% significance level and six ANN models perform better at the 5% significance level

improvement in forecasting performance. Regarding robustness, there is a slight increase (decrease in mean standard deviation of model forecasts) in the overfit training configurations.[11] Increasing the ensemble size from 50 to 100 has a smoothing effect on the ANN forecasts and results in an increased forecasting performance. This observation goes hand in hand with the residuals illustrated in Figure 9, where the overfit designs produce smoother residuals compared to their respective fit design. The observation holds for the more sophisticated networks in terms of robustness, whereas both fit and overfit designs provide sufficiently robust forecasts. As was observed for the number of neurons per hidden layer, the models in the overfit configurations tend to converge more strongly as more weight optimizations are conducted in the training process, which turns out to be slightly more robust. However, the overfit does not necessarily yield a better performance. The RMSE and MAPE show no clear tendency towards the "fit" or "overfit" training as both perform similarly well.[12] Generally, the standard deviation of the forecasts within the ensembles indicates the necessity to build ensembles as the training results in different networks. Conversely, for the SARIMAX approach, only one model is calculated for each forecasting step. The average standard deviation of all SARIMAX models is 140.03. However, this robustness measure is hardly comparable to the standard deviation of the ANN described above. While for the ANN we report an empirical standard deviation of residuals, the σ of SARIMAX is the mean theoretical standard deviation of the 37 forecast models.

To verify the statistical significance of the results, Figure 10 presents the results of a one-sided Diebold–Mariano test. The test compares two time series of residuals and indicates whether one is significantly lower than the other—that is, whether one forecast model is significantly better than the other (Diebold & Mariano, 1995). We find that the results reported in Table 3 mostly prove significant. The designs with expanding training window are significantly better than those trained with rolling training window on at least 5% significance level. The best three models in terms of MAPE and RMSE (FF_10_E_F, FF_10_E_Ov, FF_20_E_Ov) are better than the SARIMAX approach at the 1% significance level.

[11]The standard deviation amongst the 50 predictions of each model for each type is calculated for each step and then averaged over all prediction steps. In the presentation of results, it was considered that the ensemble size of the "overfit" designs is twice that of the "fit" designs. However, repeatedly sampling 50 observations from the overfit ensembles shows that the ensemble size is not decisive for the robustness measure. However, the forecasting performance decreases with reduction of the ensemble size in the overfit training strategy, particularly strong in configurations with rolling training windows.

[12]As mentioned earlier, a further sophistication of model configurations with recurrent structures and gated-recurrent units did not yield improvements compared to the networks presented in this paper. This is in line with the results for the configurations with two hidden layers compared to the one with one hidden layer and the tendency to overtraining for the more sophisticated configurations in the results presented. For more details, see the Annex.

However, the SARIMAX and naive approach compete well with the forecast performance of the ANN with rolling training window. The best ANN design in terms of MAPE and RMSE (FF_10_E_Ov) dominates all but the second-best design in terms of RMSE (FF_20_E_Ov) at 5% significance level.

To conclude, the variety of models examined in this paper offers another approach to forecasting. A solution to ally the strengths of the model classes and configurations and to balance out the shortfalls can consist of combining the different approaches. However, to build the best combination of approaches for each forecasting step, one must be aware of the strengths and weaknesses of the approaches and build a subjective market expectation. For this task, human experience is inevitable.

Finally, it is worth mentioning that the goal of this paper is to investigate approaches and configure a suitable model framework to forecast FCR prices. The presentation of the results focuses rather on the comparison of the different approaches than on the detailed discussion of single models and their coefficients' interpretation, as our goal was not to uncover the influence of the explaining variables, but to determine the performing modeling approaches and model configurations for FCR price forecasting. However, to gain more insights regarding the interdependencies and predictive power of the single exogenous variables, a detailed investigation of exemplary models from the considered approaches is an interesting direction for future research.

5 | CONCLUSION AND OUTLOOK

In this paper, we investigated approaches to forecast the price of FCR, the fastest balancing reserve that is jointly procured in weekly auctions by TSOs in Austria, Belgium, France, Germany, the Netherlands, and Switzerland. As this research scope was not formerly discussed in literature, several approaches were deployed, considering autoregressive and exogenous variables. Such a model framework has, to our knowledge, not been formerly set up or discussed.

The exogenous factors with most explanatory power are identified as the *price range* of the previous auction, the *future prices* of the German–Austrian and the French market area, the *load* in the German–Austrian and the French market area and the *planned unavailable* capacity in Germany and France. The models based on autoregressive and exogenous factors are suitable to forecast prices. Within the developed models, ANN with expanding training window yield desirable results and clearly outperform the naive forecast and the SARIMAX approach. Simple models equipped with a slight overfit and a larger ensemble size outperform the simple models that were trained aspiring to the best fit and lead to the best and most robust forecast results in the case of forecasting FCR prices. With an increase in model complexity, the positive effect of the slight overfitting strategy vanishes. Furthermore, the overall forecasting performance is not improved by more sophisticated models, as these might overtrain the relationships in the training data.

In the interpretation of these results, one must always bear in mind that econometrics and artificial intelligence approaches are only capable of drawing conclusions from data of the past. Thus changed bidding behavior by market participants or technological changes in FCR market are hardly predictable by these kinds of forecasting models. Based on assumptions (e.g., market diffusion of battery storages, market exit of conventional power plants) we could consider forecasts for the long-term FCR price development. However, this is not in the scope of this paper and needs to be addressed by future research. The main contributions of this paper are the application and comparison of statistical and neural network models to FCR price forecasting. This comprises the definition of an appropriate target variable as well as the discussion of modeling techniques and training strategies for forecasting on a scarce data basis. Finally, a discussion on the suitability and performance of simple and more sophisticated network structures for FCR price forecasting completes the contributions.

In the ongoing research, the models will be used as a basis for the formulation and optimization of bidding strategies in the European balancing reserve market. In this context, the application of SARIMAX models has the advantage that the models are open to an interpretation of the estimated coefficients, whereas the ANN approaches tend to be black boxes that yield the best results, especially in times of increased FCR price volatility, but lack interpretability. The reestimation and number of models complicate a fundamental model interpretation, as model lags, parameters and coefficients vary between the models. However, the goal in this paper is to make the forecast as accurate as possible, and reestimation increases the quality of the forecast.

Finally, the market design for FCR is in an ongoing process of change. On the one hand, the involved TSOs changed the product duration from 1 week to 1 day beginning July 2019 and intend to move to 4-hour products in the near future. This makes the consideration of forecast-based exogenous factors like wind and solar generation possible and necessary in price formation and therefore needs to be included in future studies of FCR prices. In the course of these changes, the pricing rule changed from pay-as-bid to uniform

pricing. However, the approaches developed in this study are well suited and extendable to cope with these changes and to produce reliable forecasts of FCR prices in a modified market design.

ACKNOWLEDGMENTS

This paper is a part of the work within the SINTEG project C/sells, funded by the German Federal Ministry for Economic Affairs and Energy (grant number 03SIN120).

DATA AVAILABILITY STATEMENT

Data are available on request from the authors. The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID


Emil Kraft https://orcid.org/0000-0001-9237-4488

REFERENCES

Adya, M., & Collopy, F. (1998). How effective are neural networks at forecasting and prediction? A review and evaluation. *Journal of Forecasting*, *17*(5–6), 481–495.

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales. *Journal of Retailing and Consumer Services*, *8*(3), 147–156. https://doi.org/10.1016/S0969-6989(00)00011-4

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Binner, J., Bissoondeeal, R., Elger, T., Gazely, A., & Mullineux, A. (2005). A comparison of linear forecasting models and neural networks: An application to Euro inflation and Euro Divisia. *Applied Economics*, *37*(6), 665–680. https://doi.org/10.1080/0003684052000343679

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. Cham, Switzerland: Springer. https://doi.org/10.1007/978-3-319-29854-2

Bublitz, A., Keles, D., & Fichtner, W. (2017). An analysis of the decline of electricity spot prices in Europe: Who is to blame? *Energy Policy*, *107*, 323–336. https://doi.org/10.1016/j.enpol.2017.04.034

Catalão, J. P. S., Mariano, S. J. P. S., Mendes, V. M. F., & Ferreira, L. A. F. M. (2007). Short-term electricity prices forecasting in a competitive market: A neural network approach. *Electric Power Systems Research*, *77*(10), 1297–1304. https://doi.org/10.1016/j.epsr.2006.09.022

Chatfield, C. (1996). Model uncertainty and forecast accuracy. *Journal of Forecasting*, *15*(7), 495–508.

Church, K. B., & Curram, S. P. (1996). Forecasting consumers' expenditure: A comparison between econometric and neural network models. *International Journal of Forecasting*, *12*(2), 255–267. https://doi.org/10.1016/0169-2070(95)00631-1

Diebold, F. X., & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, *13*(3), 253–263.

Dimoulkas, I., Amelin, M., Hesamzadeh, M. (2016). Forecasting balancing market prices using hidden Markov models. In 2016 13th International Conference on European Energy Market (EEM), 6–9 June 2016, Porto, Portugal.

EEX. (2019). EEX market data. Retrieved from http://www.eex.com

ENTSO-E. (2019). Transparency platform of European Network of Transmission System Operators for Electricity (ENTSO-E) Retrieved from https://transparency.entsoe.eu

Giovanelli, C., Sierla, S., Ichise, R., & Vyatkin, V. (2018). Exploiting artificial neural networks for the prediction of ancillary energy market prices. *Energies*, *11*(7), 1906–1928. https://doi.org/10.3390/en11071906

Glorot, X., Bordes, A., Bengio, Y. (2011). Deep sparse rectifier neural networks. In 14th International Conference on Artificial Intelligence and Statistics (AISTATS) 2011, Fort Lauderdale, FL.

Hwang, S., & Valls Pereira, P. L. (2006). Small sample properties of GARCH estimates and persistence. *European Journal of Finance*, *12*(6–7), 473–494. https://doi.org/10.1080/13518470500039436

Hyndman, R., & Athanasopoulos, G. (2013). *Forecasting: Principles and practice* (Print ed.). Melbourne, Australia: OTexts.

Hyndman, R., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast Package for R. *Journal of Statistical Software*, *27*(3); 1–22. https://doi.org/10.18637/jss.v027.i03

Just, S., & Weber, C. (2008). Pricing of reserves: Valuing system reserve capacity against spot prices in electricity markets. *Energy Economics*, *30*(6), 3198–3221. https://doi.org/10.1016/j.eneco.2008.05.004

Kaastra, I., & Boyd, M. (1996). Designing a neural network for forecasting financial and economic time series. *Neurocomputing*, *10*(3), 215–236. https://doi.org/10.1016/0925-2312(95)00039-9

Keles, D., Scelle, J., Paraschiv, F., & Fichtner, W. (2016). Extended forecast methods for day-ahead electricity spot prices applying artificial neural networks. *Applied Energy*, *162*, 218–230. https://doi.org/10.1016/j.apenergy.2015.09.087

Kiesel, R., & Paraschiv, F. (2017). Econometric analysis of 15-minute intraday electricity prices. *Energy Economics*, *64*, 77–90. https://doi.org/10.1016/j.eneco.2017.03.002

Kirsch, L. D., & Singh, H. (1995). Pricing ancillary electric power services. *Electricity Journal*, *8*(8), 28–36. https://doi.org/10.1016/1040-6190(95)90014-4

Klæboe, G., Eriksrud, A., Fleten, S. (2013). Benchmarking time series based forecasting models for electricity balancing market prices (Working paper). Washington, DC: George Washington University, Center of Economic Research.

Kraft, E., Keles, D., Fichtner, W. (2018). Analysis of bidding strategies in the German control reserve market. In 15th EEM, Lodz, Poland, 27–29 June 2018.

Kraft, E., Rominger, J., Mohiuddin, V., & Keles, D. (2019). Forecasting of frequency containment reserve prices using econometric and artificial intelligence approaches. In *11th International Symposium on Energy Economics*, TU Vienna (IEWT), Vienna, Austria, 13–15 February 2019.

Kwiatkowski, D., Phillips, P., Schmidt, P., & Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, *54*(1–3), 159–178. https://doi.org/10.1016/0304-4076(92)90104-Y

Lago, J., Ridder, F., & Schutter, B. (2018). Forecasting spot electricity prices: Deep learning approaches and empirical comparison

of traditional algorithms. *Applied Energy*, *221*, 386–405. https://doi.org/10.1016/j.apenergy.2018.02.069

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Liu, X. Q., Ang, B. W. & Goh, T. N. (1991): Forecasting of electricity consumption: a comparison between an econometric model and a neural network model. In: 1991 IEEE International Joint Conference on Neural Networks. Singapore: IEEE, 1254-1259 Vol. 2.

Ng, H., & Lam, K. P. (2006). How does sample size affect GARCH Models? In *9th Joint Conference on Information Sciences (JCIS)*, Taiwan, ROC, 08–11 October 2006. Paris, France: Atlantis Press.

Ocker, F., Ehrhart, K.-M., & Belica, M. (2018). Harmonization of the European balancing power auction: A game-theoretical and empirical investigation. *Energy Economics*, *73C*, 194–211.

Ocker, F.; Ehrhart, K.-M. (2017). The "German Paradox" in the balancing power markets. In: *Renewable and Sustainable Energy Reviews*, *67*, S. 892–898. https://doi.org/10.1016/j.rser.2016.09.040

Oksuz, I., & Ugurlu, U. (2019). Neural network based model comparison for intraday electricity price forecasting. *Energies*, *12*(23), 4557–4571. https://doi.org/10.3390/en12234557

Olson, D., & Mossman, C. (2003). Neural network forecasts of Canadian stock returns using accounting ratios. *International Journal of Forecasting*, *19*(3), 453–465. https://doi.org/10.1016/S0169-2070(02)00058-4

Olsson, M., & Söder, L. (2008). Modeling real-time balancing power market prices using combined SARIMA and Markov processes. *IEEE Transactions on Power Apparatus and Systems*, *23*(2), 443–450. https://doi.org/10.1109/TPWRS.2008.920046

Qi, M., & Zhang, G. P. (2001). An investigation of model selection criteria for neural network time series forecasting. *European Journal of Operational Research*, *132*(3), 666–680. https://doi.org/10.1016/S0377-2217(00)00171-5

Ramachandran, P., Barret, Z., & Quoc, L. V. (2017). Searching for activation functions. arXiv: 1710.05941, https://arxiv.org/pdf/1710.05941v2.

regelleistung.net. (2019). Internetplattform zur Vergabe von Regelleistung [Internet platform for balancing power procurement]. Retrieved from https://www.regelleistung.net

Tkacz, G. (2001). Neural network forecasting of Canadian GDP growth. *International Journal of Forecasting*, *17*(1), 57–69. https://doi.org/10.1016/S0169-2070%2800%2900063-7

Ugurlu, U., Oksuz, I., & Tas, O. (2018). Electricity price forecasting using recurrent neural networks. *Energies*, *11*(5), 1255–1276. https://doi.org/10.3390/en11051255

Wang, P., Zareipour, H., & Rosehart, W. (2014). Descriptive models for reserve and regulation prices in competitive electricity markets. *IEEE Transactions on Smart Grid*, *5*(1), 471–479. https://doi.org/10.1109/TSG.2013.2279890

Weron, R. (2014). Electricity price forecasting. A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, *30*(4), 1030–1081. https://doi.org/10.1016/j.ijforecast.2014.08.008

Yao, J., & Tan, C. L. (2000). A case study on using neural networks to perform technical forecasting of forex. *Neurocomputing*, *34*(1–4), 79–98. https://doi.org/10.1016/S0925-2312(00)00300-3

AUTHOR BIOGRAPHIES

Emil Kraft studied at École polytechnique fédérale de Lausanne (EPFL) and Karlsruhe Institute of Technology (KIT) and graduated as Industrial Engineer in 2017. Currently, he is research associate and PhD candidate in the research group "Energy Markets and Energy System Analysis" at the Chair of Energy Econmics at the Institute for Industrial Production. His research focuses on the analysis and modelling of balancing energy markets and decision taking under uncertainty in the electricity markets context.

Dr. Dogan Keles graduated as Industrial Engineer from Karlsruhe Institute of Technology (KIT) in 2006 and received his PhD (Dr. rer. pol.) in 2013. During his work at the KIT he analysed uncertainties in energy markets and developed methods to evaluate energy investments. During his Senior Research Fellowship at Durham University in 2019, he carried out research on the effect of RES on electricity markets and designing systems with large shares of renewables. Currently, Dogan Keles is head of the research group "Energy Markets and Energy System Analysis" at the Chair of Energy Econmics.

Prof. Dr. Wolf Fichtner studied Industrial Engineering and Management at the University of Karlsruhe (TH) from 1988 on. In 1998 he received his PhD (Dr. rer. pol.). After his habilitation in 2004, he initially worked as a university lecturer at the Institute for Industrial Production before joining EnBW Energie Baden-Württemberg AG as a project manager. In 2005, he accepted a call to the Brandenburg University of Technology in Cottbus, where he was head of the Department of Energy Economics until 2008. Since the end of 2008, he has been head of the Chair of Energy Economics at the Institute for Industrial Production at Karlsruhe Institute of Technology and head of the French-German Institute for Environmental Research (DFIU). Since 2012, he has been Director at the Karlsruhe Service Research Institute.

How to cite this article: Kraft E, Keles D, Fichtner W. Modeling of frequency containment reserve prices with econometrics and artificial intelligence. *Journal of Forecasting*. 2020;1–19. https://doi.org/10.1002/for.2693

APPENDIX A

Here, supplementary illustrations of the results are presented as well as the hyperparameters, training strategies, and results of the network structures with GRU that were mentioned but not reported in Sections 3 and 4.

Figure A shows the histograms of residuals of FCR price forecasts in the test period. Figures B and C show the FCR price forecasts and residuals in the test period that were not shown in Figures 7 and 9 but reported in Table 3.

Tables A and B show the experimental design deployed for the GRU neural networks and the forecasting results. Hereby, one design consists of the combination of the hyperparameters *number of hidden layers*, *number of neurons* per hidden layer, and the *training strategy* defined by the training data and the training hyperparameters that are provided in Table A. Table B provides, analogously to Table 3, the performance indicators RMSE, MAPE, and DAC, and the robustness measure σ of the model forecasts for the GRU networks. Regarding forecasting performance, no improvement to the feedforward networks can be observed. However, regarding robustness, the standard deviations are generally smaller, which indicates model training is converging more strongly compared to the feedforward networks. To conclude, in our case the models with GRU lead to more robust forecasts around less accurate estimates.

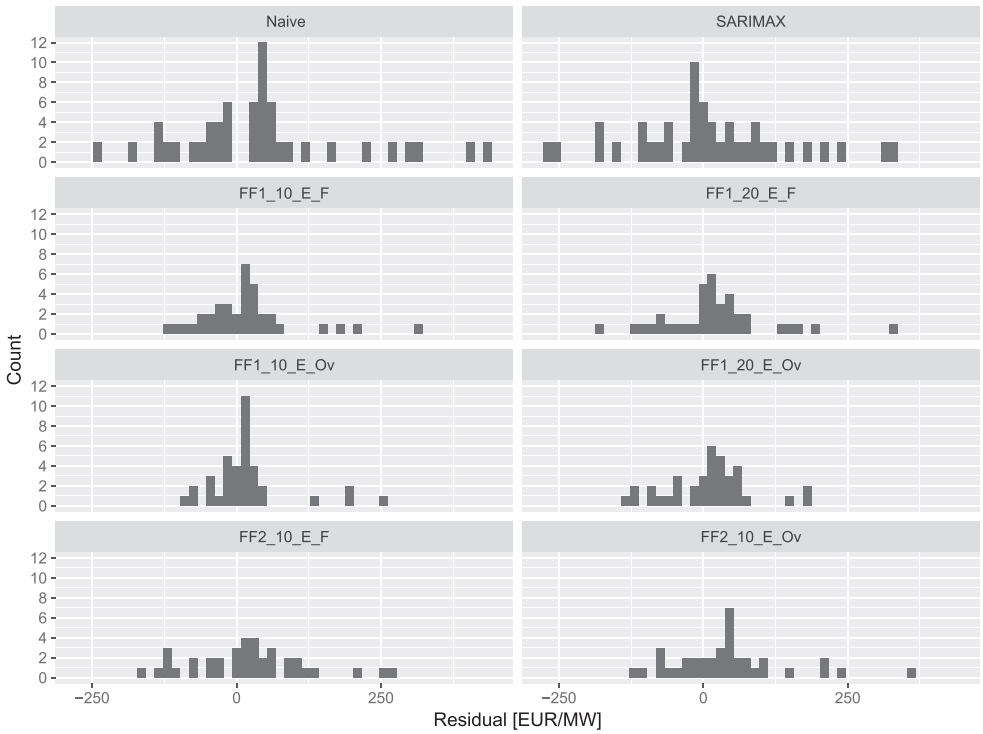

FIGURE A Histograms of residuals of FCR price forecasts in test period 2018:Q1–Q3. It can be observed that the ANN residuals follow a relatively symmetric distribution around zero. The better models have a higher count in the bins closer to zero. To derive more insights regarding the residuals' distribution, more observations would be required

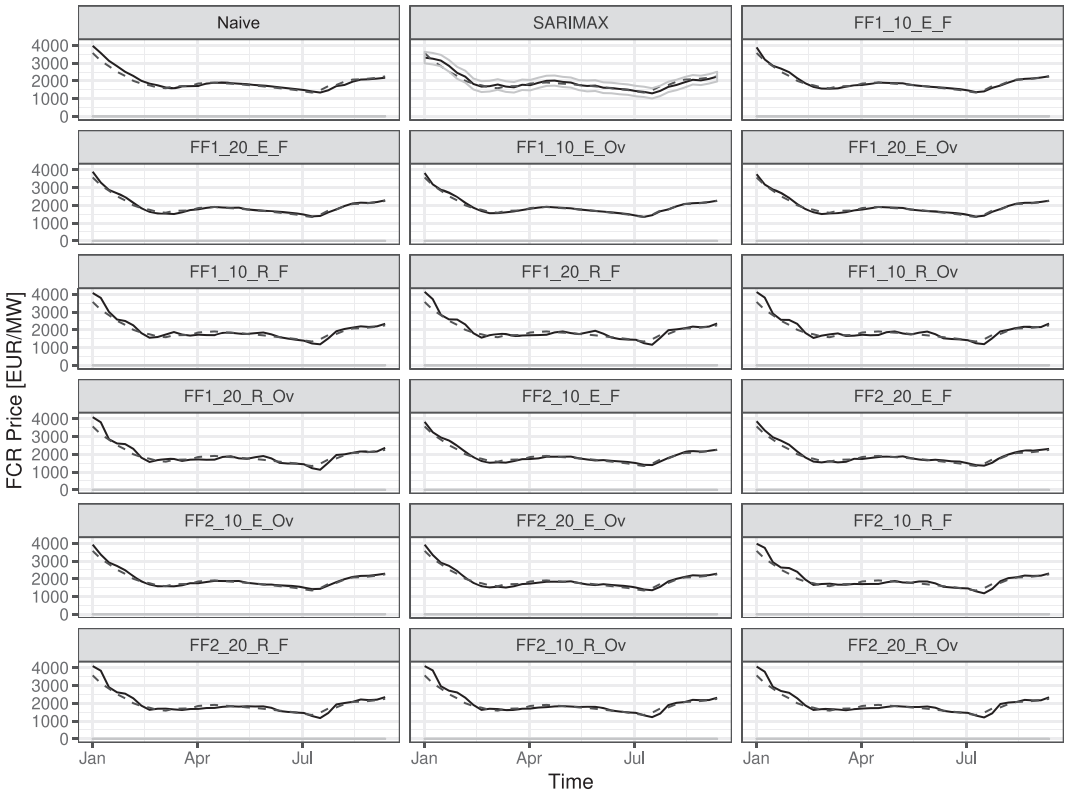

FIGURE B FCR price forecasts in test period 2018:Q1–Q3 (original FCR price data from regelleistung.net, 2019). Solid lines are the forecasted FCR prices, and dashed lines represent the realized FCR price. In addition to the choice of models presented in Section 4, for completeness all deployed model designs reported in the results are shown, indicating a suitable fit for all designs

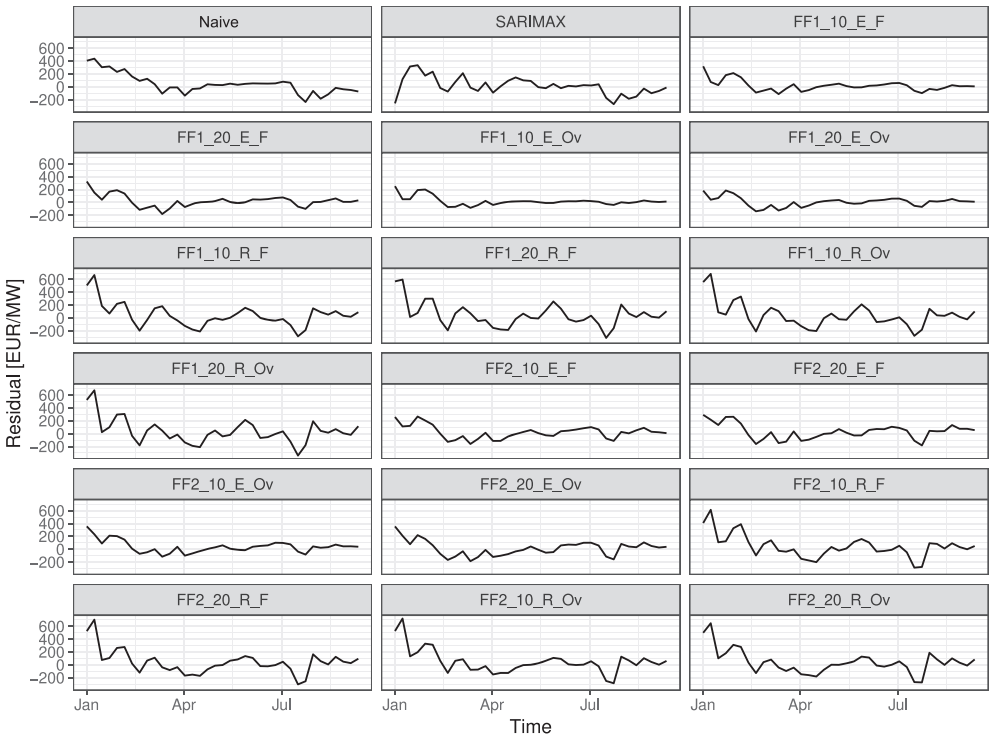

FIGURE C Residuals of FCR price forecasts in test period 2018:Q1–Q3 (original FCR price data from regelleistung.net, 2019). For completeness and supplementary to the choice of models presented in Section 4, all deployed model designs reported in the results are shown. The residuals of similar model configurations show similar residual shapes

TABLE A Experimental design for GRU neural networks

Factor	Level reference	Level variation
Network configuration		
Number of hidden layers	GRU1 (1 hidden layer)	GRU2 (2 hidden layers)
Number of neurons per hidden layer	10	20
Training strategy		
Training data	E (expanding window)	R (rolling window)
Training hyperparameters	F (fit): number of batches = 1, number of epochs = 30, iterations per epoch = 20, ensemble size = 50	Ov (Overfit): number of batches = 1, number epochs = 50, iterations per epoch = 30, ensemble size = 100

TABLE B Root mean square error (RMSE), mean absolute percentage error (MAPE), directional accuracy (DAC), and mean standard deviation (σ) of the model forecasts with GRU networks

Design	RMSE	MAPE	DAC	σ
GRU1_10_E_F	124.47	4.65%	91.70%	50.58
GRU1_20_E_F	190.71	5.49%	86.10%	170.09
GRU1_10_E_Ov	142.55	4.81%	88.90%	68.92
GRU1_20_E_Ov	174.25	6.03%	83.30%	69.37
GRU1_10_R_F	166.00	7.99%	86.10%	62.56
GRU1_20_R_F	205.11	10.24%	77.80%	57.60
GRU1_10_R_Ov	197.59	9.35%	80.60%	74.79
GRU1_20_R_Ov	220.20	10.57%	77.80%	99.65
GRU2_10_E_F	151.74	5.10%	88.90%	44.48
GRU2_20_E_F	174.61	5.66%	86.10%	70.13
GRU2_10_E_Ov	216.30	9.52%	80.60%	80.16
GRU2_20_E_Ov	205.19	6.71%	77.80%	195.88
GRU2_10_R_F	196.78	8.52%	80.60%	55.06
GRU2_20_R_F	224.23	10.05%	77.80%	55.64
GRU2_10_R_Ov	217.63	9.68%	77.80%	79.12
GRU2_20_R_Ov	224.86	10.12%	77.80%	83.73

Note. No improvement compared to the feedforward networks is achieved. The standard deviations are generally smaller, which indicates a more robust model training, but RMSE, MAPE, and DAC indicated no better forecasting performance.